Intro to Machine Learning Coursework 1 Report - Decision Trees

Musawer Hussain, Vladimir Volgin, Oleg Tretieu

3rd November 2023

1 Introduction

This is the final report of the first Introduction to Machine Learning coursework where we have implemented a decision tree algorithm and use it to determine one of the indoor locations based on WIFI signal strengths collected from a mobile phone. We were given two datasets, a clean dataset and noisy dataset and we have trained decision trees on each of them separately using 10-fold cross-validation.

In this report we will first show the output of one of our decision trees formed, followed by listing out the evaluation metrics we have calculated for decision trees trained on both clean and noisy datasets before finally providing a short analysis of both the results and the difference in the two datasets.

2 Image of the Decision Tree Formed

The following is the Decision Tree that was formed as a result of training on the full clean-dataset. This is the output of the visualisation function.

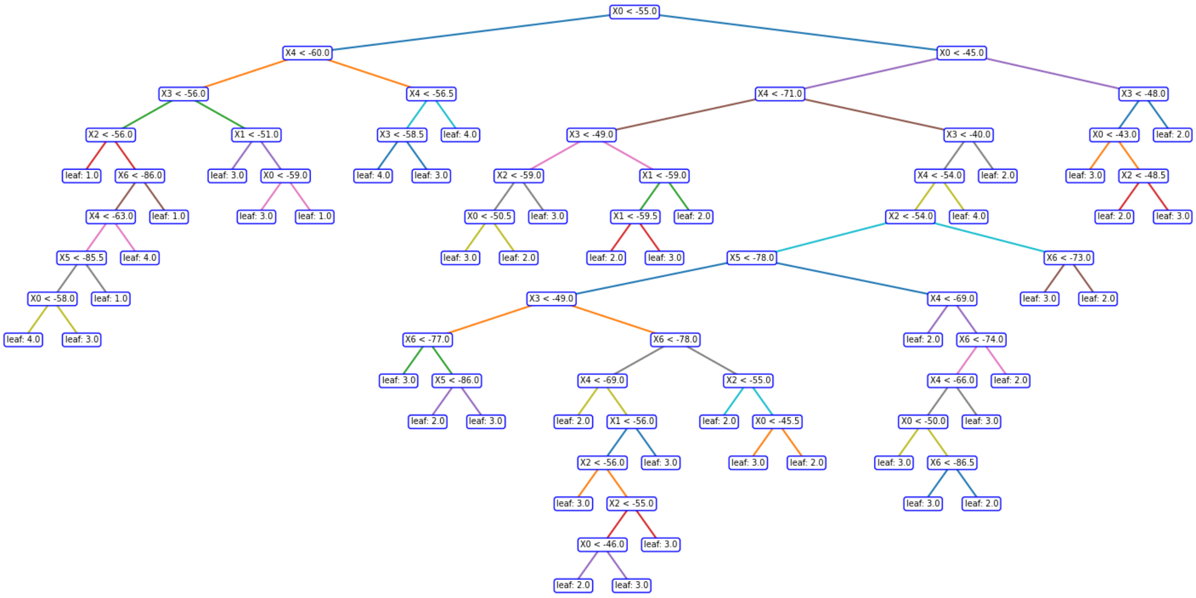

Figure 1: Output of the decision tree visualisation function on the clean dataset.

3 Evaluation Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

3.1 Clean data set

	Room 1 predicted	Room 2 predicted	Room 3 predicted	Room 4 predicted
Room 1 actual	496	0	2	2
Room 2 actual	0	478	22	0
Room 3 actual	1	18	479	2
Room 4 actual	3	0	0	497

Table 1: Confusion matrix for clean dataset

Accuracy for each fold:	**[0.98 0.98 0.955 0.98 0.975 0.98 0.97 0.96 0.975 0.97]**
Average Accuracy:	**0.975**
Recall:	
	Room 1: 0.992
	Room 2: 0.956
	Room 3: 0.958
	Room 4: 0.994
Precision:	
	Room 1: 0.992
	Room 2: 0.964
	Room 3: 0.952
	Room 4: 0.992
F1-measures:	
	Room 1: 0.992
	Room 2: 0.960
	Room 3: 0.955
	Room 4: 0.993
Macro-averaged Recall:	**0.972**
Macro-averaged Precision:	**0.972**
Macro-averaged F1-Measure:	**0.972**

3.2 Noisy data set

	Room 1 predicted	Room 2 predicted	Room 3 predicted	Room 4 predicted
Room 1 actual	383	33	38	36
Room 2 actual	33	387	46	31
Room 3 actual	24	47	412	32
Room 4 actual	40	27	34	397

Table 2: Confusion matrix for noisy dataset

	Room 1 predicted	Room 2 predicted	Room 3 predicted	Room 4 predicted
Room 1 actual	0.782	0.067	0.0776	0.0734
Room 2 actual	0.0664	0.779	0.0926	0.0624
Room 3 actual	0.0466	0.0913	0.800	0.0621
Room 4 actual	0.0803	0.0542	0.0683	0.797

Table 3: Normalized Confusion matrix for noisy dataset

Due to the noisy dataset being slightly unbalanced we decided to compute the normalized confusion matrix as well. However due to the unbalance being very small the evaluation metrics were similar across both matrices and so the metrics from the normalized matrix have been omitted.

Accuracy for each fold:	**[0.775 0.785 0.795 0.77 0.765 0.795 0.84 0.785 0.805 0.78]**
Average Accuracy:	**0.790**
Recall:	
	Room 1: 0.782
	Room 2: 0.779
	Room 3: 0.800
	Room 4: 0.797
Precision:	
	Room 1: 0.798
	Room 2: 0.783
	Room 3: 0.777
	Room 4: 0.800
F1-measures:	
	Room 1: 0.790
	Room 2: 0.781
	Room 3: 0.789
	Room 4: 0.800
Macro-averaged Recall:	**0.789**
Macro-averaged Precision:	**0.790**
Macro-averaged F1-Measure:	**0.790**

4 Analysis

4.1 Results Analysis

When using the clean data, the vast majority of the rooms are identified correctly (shown by the high accuracy). We see from the confusion matrix that the rooms which were most often confused with eachother are Room 2 and Room 3. With the noisy data, while most rooms are still correctly recognised, there is a significant amount of incorrect classifications. Most notably: Room 1 often gets misclassified as Room 2, Room 4 as Room 1, and Room 2 as Room 3.

4.2 Dataset Differences

The performance on the noisy dataset is significantly worse than on the clean dataset, which we can see by comparing the evaluation metrics above. This could be caused by the decision tree overfitting when training on the noisy data, effectively "learning the noise", which leads to worse generalisation. Furthermore, the noisy dataset is slightly unbalanced, as Room 3 has 515 samples, whereas Room 1 only has 490, so it is necessary to look at the normalized confusion matrix.